Visualization as Integration of Heterogeneous Processes

Xiaoyu Wang[a], Wenwen Dou[a],William Ribarksy[a],Remco Chang[a]

[a]Visualization Center, University of North Carolina at Charlotte, Charlotte, USA

ABSTRACT

In the information age today, we are experiencing an explosion of data and information from a variety of sources unlike anything that the world has seen before. While technology has advanced to keep up with the collection and storage of data, what we lack now is the ability to analyze and understand the meaning behind the data. Traditionally, data mining and data management techniques require the data to be uniform such that a single process can search for knowledge within the data. However, in analysis of complex tasks where knowledge and information need to be pieced together from different sources of data, a new paradigm is required. In this paper, we present a framework of using visual analytical approaches to integrate multiple heterogeneous processes that can each analyze a specific type of data. Under this framework, stand-alone software solutions can focus on specific aspects of the problem based on domain-specific techniques. The framework serves as a visual repository for all the information and knowledge discovered by each individual process, and allows the user to interactively perform sense-making analysis to form a cohesive and comprehensive understanding of the problem at hand. We demonstrate the effectiveness of this framework by applying it to inspecting bridge conditions that utilizes data sources from 2D imagery, 3D LiDAR, and multi-dimensional data based on bridge reports.

Keywords: Heterogeneous Data, Heterogeneous Processes, Visual Analytics, Integration, Sense-Making

1. INTRODUCTION

With people monitoring and collecting data from various sources, a massive amount of heterogeneous data have been created. Since those sources are not centrally managed, the data do not follow existing or known formats and often contain conflicting information and knowledge that need special interpretations and representations. Although certain heterogeneous data could be presented as unified dataset using advanced data managing techniques, it is quite expensive and time consuming to transform and unify them. More importantly, it is probable that the transformation of data would result in losses of certain original knowledge.

Despite the difficulty in unifying data, the use of heterogeneous data is crucial for investigative tasks and sense-making processes. This is because complex problems, such as those faced by the Department of Homeland Security, often require the examination of disparate data sources for the underlying trends or patterns to become apparent. Due to the limited information carried by a single dataset, utilizing heterogeneous data allows user to gather more comprehensive information that could lead to a better understanding of the tasks and make corresponding decisions.

However, the need for examining multiple sources does not imply unifying them into a coherent dataset. In most real life scenarios, analysts have different tools for foraging different types of data and analyzing them. Unfortunately, making sense of the results of the analysis from using these tools is often difficult and confusing without any tools to facilitate the process. In this paper, we present our visual analytics framework that uses visualizations to integrate heterogeneous processes. Instead of directly managing low-level datasets, our framework focuses on managing the processes that examine the datasets and collects the resulting analysis of each process into a coherent comprehensive picture. Each process hosts certain domain viewpoints and provides the user with a semantic representation of a specific dataset. Our visual analytics system does not require all processes to be integrated or built by the same system designers. To enable the information sharing between different processes, our system adopts a communication protocol that is agreed upon by all processes that could be as simple as passing certain predefined ID numbers or as complex as semantic contents. Each process could

Further author information: (Send correspondence to Xiaoyu Wang.)
Xiaoyu Wang: E-mail: xwang25@uncc.edu, Telephone: 1 704 687 8641

choose to be synchronous or asynchronous to the changes from other processes. With all the integrated processes, our framework provides the user a comprehensive visual analytics environment to facilitate reasoning tasks.[?]

In order to prove the feasibility of this framework, we built a prototype system that assists bridge managers at North Carolina Department of Transportation to perform bridge inspection and analysis. The system integrates several heterogeneous processes, such as an LIDAR display, an image processing software and an ontological knowledge structure.

In the remainder of this paper, Section 2 is structured to present the relationship between task and process. We give an overview of related work in Section 3. In Section 4, we present details of our framework, followed by presenting and discussing application in Section 5. We provide our conclusion and future works in Section 6.

2. TASKS AND PROCESSES

Solving a complex task can often be accomplished by solving multiple individual sub-tasks. These sub-tasks can further be categorized in a hierarchical fashion such that the goals and the required actions are specified.[?]

In this paper, we specifically focus on analytical and investigative tasks, and we refer to the actions taken towards solving these tasks as processes, which could include the use of individual computer programs or a set of computer analysis steps. For example, in the counter terrorism field, a task could be the understanding a group of terrorists' characteristics and predict their future activities. In this example, individual processes could include the search for when their last appearances was (by examining temporal information), where was their last target (through the use of geospatial information) and how did they attacked (through analyzing news and reports).[?]

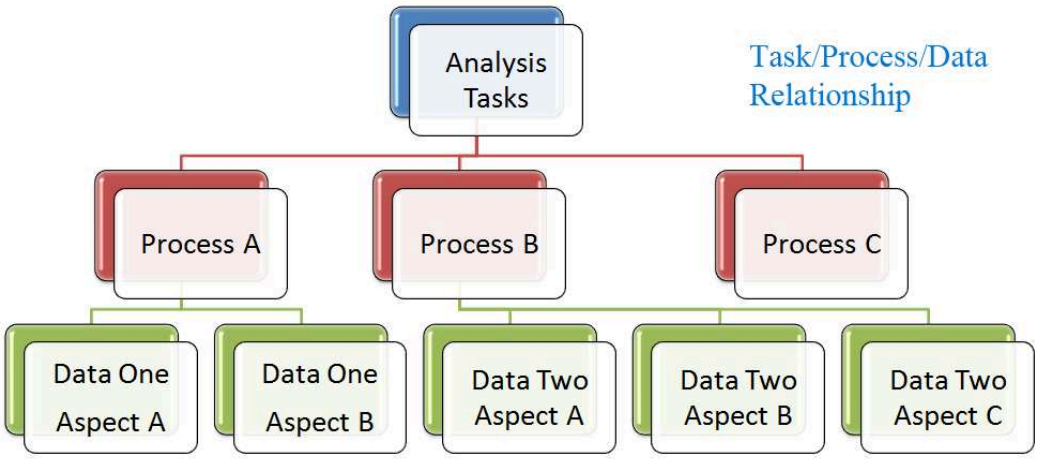

Figure 1. The relationship between Tasks, Processes and Data

An individual process could be designed to analyze a certain type of data. In modern computing, computer processes have become more specialized in analyzing those preparatory data. For instance, advanced image analysis tools are used to analyze terrain changes[?] or longest common string techniques used for searching and matching gene sequences.[?] While these advanced processes provide their user deep understanding about a specific domain, when applied to an investigative task, it is often difficult to piece them together and gain comprehensive understanding on all facets of the data. These problems are like the challenges faced by criminologists where the investigators need to manually piece together evidence from various analysis results such as blood samples, DNA tests or even bone structures. Although each of these results could be effectively calculated by individual computer process, there is no simple way to efficiently depict every one of them as the amount of information could be overwhelming. Hence, these manual investigations would take tremendous amount of efforts that could still sometime result in incomplete analysis, as shown in Figure **??**.

Although the most preferable method to obtain comprehensive understandings about the data is performing these processes on a unified database, it is not always convenient to acquire or retain such database. For one, it is hard to come up with such unified data schematic. Different data source is collected disparately using individual data schematics. With the vast diversities in today's data, it is less common to have experts who could understand or standardize all these data sources to create such unified database. Secondly, integrating data together has the potential risk of losing certain contexts when analysis of the preparatory data could only make sense within those contexts. For example, the characteristics of a terrorist group are frequently tightly related to its cultural background and the political climate at the time. If analysis on this group is performed without considering these factors, the result could be quite misleading and potentially unreliable. Lastly, even when a uniform dataset is available and the context of the analysis preserved, the entire approach of integrating heterogeneous data sources requires tremendous amount of effort. Given the complex information collected today, this is a really expensive approach without much guaranteed success.

Therefore, in order to accomplish such analysis tasks, it makes more sense to focus on how to integrate different processes, but not on the integration of data. Integrating processes that are specialized in different fields and share their embedded knowledge and results could not only avoid losing analysis context, but also more effectively depict information of the different facets.

3. RELATED WORK

3.1 Data Integration

While technology has advanced to keep up with the collection and storage of data, we are still limited on the ability to analyze and understand the meaning behind the data. Traditionally, data mining and data management techniques require the data to be uniform such that a single process can search for knowledge within the data[?]

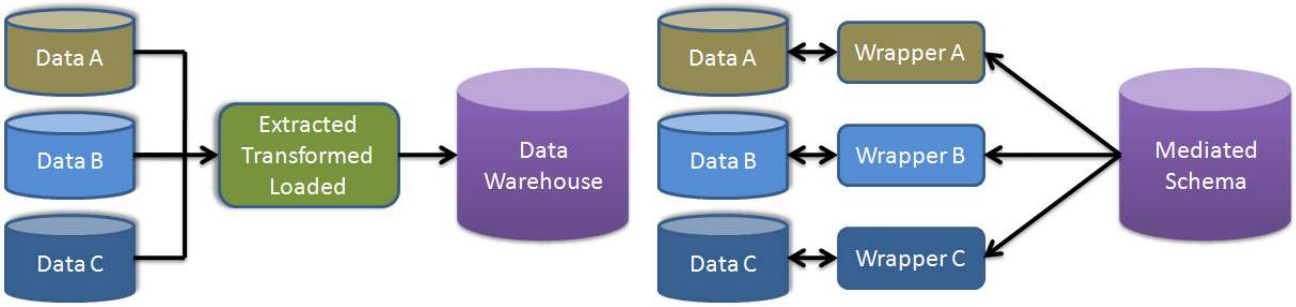

Figure 2. Two Types of typical approaches for Data Integration. Left: (A) A simple data warehouse. The information from the source databases is extracted, transformed then loaded into the data warehouse. Right: (B) A simple data integration solution through mediated schema

Data warehousing, as shown in Figure **??**(A), is an popular approach in achieving such integrated database. Although this can be perceived architecturally as a tightly coupled approach, updating the data sources has the potential of causing an expensive cascading effect. In most cases, after the original data source is updated, the warehouse will have to be updated through an extraction, transformation and loading (ETL) process.[?] Furthermore, it is also difficult to construct data warehouses when you only have a query interface to the data sources and no access to the full data.

To address this issue, the recent trend in data integration tend to loosen the coupling between data. As shown in Figure **??**(B), the integration idea is to provide a uniform query interface over a mediated schema that allows passing and sharing transformed queries. Although this is more flexible than the traditional ETL approaches, it still raises the issue known as the Semantic Integration Problem. This problem is not about how to structure the architecture of the integration, but how to resolve semantic conflicts between heterogeneous data sources. A common strategy for the resolution of such problems is the use of ontologies which explicitly defined schema terms and thus help to resolve semantic conflicts, as mentioned in.[?]

3.2 Visualizations of Data Integration

Even though the aforementioned two data integration processes could eventually merge different heterogenous data, the user still does not have the ability to analyze and understand the meaning behind the data. Since each individual data sources represent different facets of the information, merely examine a data source at a time would lead to an incomplete understanding.

Researchers in the visualization community has spent a great deal of effort in searching for a good way to visually integrate such heterogeneous data sources. Keim[?] proposed a theoretical classification of information visualization and visual data mining techniques that is based on the data type to be visualized, the visualization technique, and the interaction and distortion techniques. In practice, Callahan et al[?] proposed a visualization system, VisTrail, to perform dynamic data management. In addition, through decoupling the schema of the underlying data from the specification of a visualization, Cammarano et al[?] presented an advanced technique to automatically create visualizations to represent different data aspects.

Although these visualization approaches could assist the user in managing the data sources and depicting them from different aspects, they still restrain the user's exploration in a visualization environment. As mentioned in the previous section, analysis tasks often require a combination of analysis processes from different fields. Therefore, instead of only analyzing the data using visualization, analysts can more readily utilize the results from different analysis tools. Hence, we propose a framework to use visualization as an integration of different reasoning processes.

4. THE FRAMEWORK

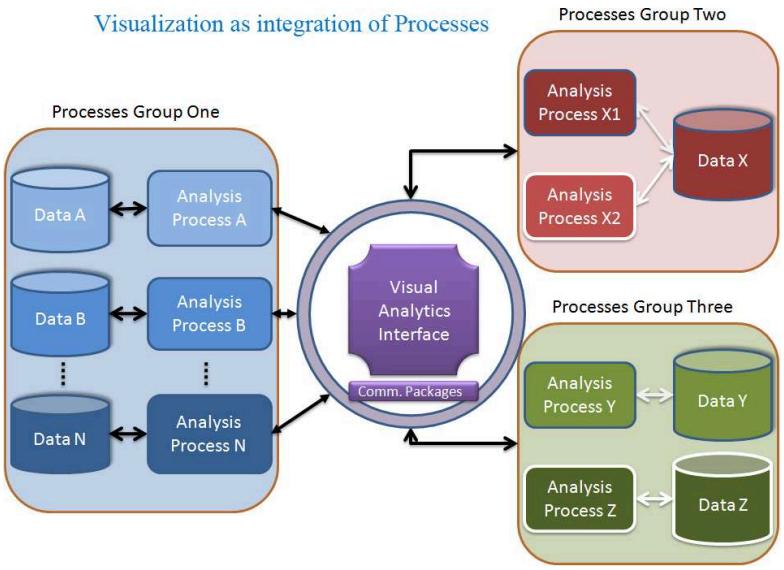

Figure 3. The Framework Overview

Since the cost of creating a unified data source could be tremendous, we propose that reviewing and analyzing multiple data sources does not imply the need to unify them into a coherent dataset. In most analysis tasks, analysts with different expertise often have specialized tools for foraging those different types of data. The results from those independent data analysis play an important role in solving the entire analysis task. Unfortunately, without certain facilitating tools or guidelines, it is often difficult and confusing to utilize these results to acquire a comprehensive understanding of the problem domain.

We present our visual analytics framework that provides such facilitation to analysts by integrating multiple heterogeneous processes. In our framework, we focus on collecting and sharing results from different analysis processes rather than directly managing the low-level datasets. Since each process is incorporated with domain viewpoints and knowledge, they provide the user with semantic representation of specific datasets. Our framework

dynamically collects and utilizes the results generated by these processes and represent them to the user in an interactive manner such that an analysts can piece them into a coherent comprehensive view, as shown in Figure **??**.

Visualization in our framework plays a key role as an intuitive integration interface that represents and manages different analysis processes. Giving its power of providing visually representing of the abstracted information[?],[?] visualization not only provides a graphical interface to switch and invoke individual process, but it also bridges and communicates to different process to retain a continuous analysis context. Analysts now could have comprehensive understanding of their analysis task and come up with accurate solutions.

Using the visualization as an integration point also enables the analysts to interactively switch between individual processes. In our framework, since each process is coordinated with others through the visual interface, analysts could easily select and change to other process. This not only enables the effective analysis of different datasets, but also reduce the confusion caused by switching context.

On a design level, our visual analytics framework allows all integrated processes to be built independently. Since individual domain experts and engineers could follow their own guidelines, our framework could significantly increase their design efficiency and provide an intuitive way to sharing analysis results.

Structured as a graph, our framework currently adopts standard communication packages to enable the information sharing between individual processes. The main visual management interface is the root node that collects and manages those communication packages, while individual processes are the child nodes that connect to the interface. The complexities of these communication packages vary for different analysis tasks. For example, it could contain certain predefined features that all the individual process would have, like terrorist group names,[?] bridge number or DNA sequence IDs.[?] In addition to these lead features, the packages could also include detail parameters that further provide more specific information. By sharing this information among different processes and allowing them to perform individual analysis, our framework eventually creates a comprehensive decision making and reasoning environment.

Since not all of the analysis process depends on others, our framework implements a flexible mechanism to passing communication packages between the processes. Each process is authorized with the choice to be synchronized or unsynchronized to the changes from other processes. This enables analyst to examine certain information in a specific context without the disturbance from the other irrelevant information. In this way, a bridge engineer could focus on examining the cracks on some pavements, without the need to deal with updated results for the main bridge structures.[?]

With all the integrated processes, our framework provides user a comprehensive visual analytics environment to facilitate the reasoning tasks.

5. APPLICATION

There are lots of fields that have such complex data integration and analysis problems, like the counter-terrorism action in DHS or the infrastructure management in Department of Transportation. We had a great opportunity to collaborate with North Carolina Department of Transportation to design an integrated system for maintaining bridge infrastructure. Since the bridge maintenance requires deep understanding of various data sources, appropriate bridge analysis becomes a complex data integration and analysis problem that is well-suited for our framework. For example, the data sources need to be reviewed include bridge inspection imageries, the LIDAR scan data, the sensor data, and etc. Although there are tools to examine most of this data, it is quite challenging for the bridge engineers to comprehensively understand them all and make accurate maintenance decisions, as shown in Figure **??**.

5.1 System Design

Following the aforementioned in the framework, we designed a visual analytics system that could integrate those specialize bridge analysis programs together. Our system not only could help bridge engineers to collect information from different analysis results but it also facilitate their decision making tasks. In our system, we have three major components, the interactive visual analytics interface, the ontological knowledge structure and

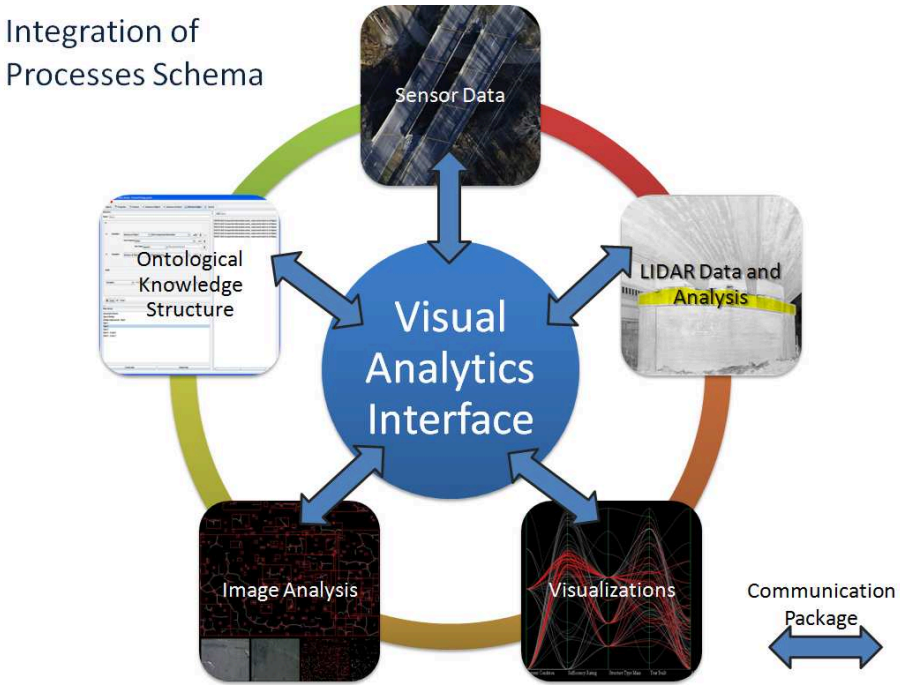

Figure 4. System Overview: (A) The Visualizations includes several highly interactive visual representations for depicting information from different data aspects. (B) The Ontological Knowledge provides the guidelines for corresponding bridge inspections domain information (C) Analysis tools, including image analysis tool for analyzing pavement, LIDAR analysis tool for inspecting bridge structure and Remote sensing tools for examining overall conditions. (D) The communication packages that is shared among these integrated components.

the peripheral analysis tools. Each component representing a set of processes and it is designed by individual engineers and collaborators for this project.

While the visual analytics interface serves as an integration of other components in this system, as shown in Figure **??**, it also provides a highly interactive data exploration environment. The ontological knowledge structure preserves and provides corresponding bridge inspections domain information. Lastly, the set of peripheral analysis tools includes a LIDAR analysis tool and a pavement analysis tool provide specific domain results for different bridge components. With all these components together, our system could provide comprehensive understandings about those bridges.

Since in our system the main featured objects are clearly the bridges and its components, the communication package are designed around depicting these features. More specifically, the package is consisted of Bridge Name, Bridge Component Name and Component Specific Parameters. These packages contain most aspects of those objects and the communications between each component are clear and feasible. Utilizing the most relevant package information, each integrated component performs it specified analysis and interactively share the results with the other components and the user.

As an integration and management of different processes, the visualization interface assists the user to depict different analysis results and invokes certain process during their data exploration. The visualization component adopts a multiple-window approach to automatically coordinate all these processes in a manner that the change in one process window will affect the other ones when the views are appropriately coordinated. This allows our system to keep the analyst within their previous reasoning context and reduce the cognitive cost of switching to different analysis results.

By communicating information between the different components, our system can now provide bridge engineers not only the ability to freely explore their preparatory data, but also facilitate them with their decision-making tasks.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose our framework to use a visual analytics interface as an integration of heterogeneous processes, each of which can analyze a specific type of data. Using this framework, we can reduce the problems introduced by traditional data mining techniques and support comprehensive analysis of the data. Under this framework, stand-alone software solutions can focus on specific aspects of the problem based on domain-specific techniques. Our framework serves as a visual management repository for all the information and knowledge discovered by each individual process, and allows the user to interactively analyze in the processes in a cohesive and comprehensive manner.

To demonstrate the feasible of our system, we collaborated with NCDOT to present an interactive visual analytics system for conducting bridge inspections. In this project, multiple data sources have been utilized by individual programs to reveal different bridge aspects, such as 2D imagery, 3D LiDAR, and multi-dimensional data based on bridge reports. The visual analytics interface automatically coordinates different programs and share communication packages between them to provide the user a complete analysis of the data.

In theory, this framework is still at a preliminary stage; there are lots of details that could be improved. For example, how much communication is necessary for two processes to fully share their analysis processes? And can the processes be aggregated or integrated into higher level analytical tasks? We also need further investigate the different relationships between task and processes, and identify more comprehensive principles for furthering this visual analytics research.

In practical, although the application we presented to NCDOT has received positive feedback, we are still a few steps away from real deployment for bridge engineers to use in their everyday life. Further investigations in this framework are necessary to understand the limitations of this framework, and how we could overcome them from both the theoretical as well as the practical perspectives.

ACKNOWLEDGMENTS

This project is supported by grant number DTOS59-07-H-0005 from the United States Department of Transportation (USDOT), Research and Innovative Technology Administration (RITA). The views, opinions, findings and conclusions reflected in this presentation or publication are the responsibility of the authors or presenters only and do not represent the official policy or position of the USDOT, RITA, or any State or other entity. The authors also would like to acknowledge the guidance and contributions of Mr. Caesar Singh, the Program Manager at USDOT; and the technical assistance of Dr. Moy Biswas of the North Carolina DOT (NCDOT), Mr. Garland Haywoord of NCDOT Division 10, and Mr. Jimmy Rhyne of Charlotte DOT.